韓國軍事學論集 第77輯 第2卷 2021年 6月 DOI : http://doi.org/10.31066/kjmas.2021.77.2.016

국방분야 빅데이터 분석을 위한 자연어 사전 구축

Building a Natural Language Processing Dictionary for Analysing Military Areas' Bigdata

홍 힘 찬(Himchan Hong)[1)]
31사단 503여단 2대대 7중대장

ABSTRACT

Analysis of unstructured texts on the national defense is a useful tool for various occasions. However, using an open-source morphological analysis model for studying big data in the military area can result in an unintended bias because the model has no prior understanding of military terminology, leading to misinterpretation of terms related to the defense issue. Therefore this study suggests a natural language dictionary for the national defense field, which can be efficient in minimizing the bias during the analysis of natural language associated with the field. In this study, five online dictionaries related to military and associated terms were collected to set a new natural language dictionary of 20,875 words. After training an analysis model with the new dictionary, the identification rate of the model improves by about 73%. Moreover, when comparing the identification rate of the trained model and the five other models with no training, the rate of the trained model was 6%p ~ 88%p higher than that of the other models. This result shows that establishing a natural language dictionary of the national defense and applying it to the analysis of unstructured text on the issue would lower the possibility of bias in the analysis result.

Key words : Morphological Analyzer, Bigdata, Natural Language Process, Dictionary of military and associated terminology
주 제 어 : 형태소 분석, 빅데이터, 자연어 처리, 국방사전

1) 주저자 : honghimchan@gmail.com

I. 서 론

1. 연구배경 및 목적

‘적을 알고 나를 알면 백번 싸워도 위태로움이 없다(知彼知己 百戰不殆)’는 손자병법의 구절과 같이 지식(知)의 활용은 동서고금을 막론하고 전쟁의 승패를 좌우하는 중요한 요소 중 하나이다. 지식에 대한 이러한 관점으로 볼 때 국방 분야에서 빅데이터에 대한 관심은 갑작스럽게 증가한 것이 아닌 이전부터 존재해왔던 것임을 알 수 있다. 따라서 빅데이터에 대한 군의 새로운 관점과 정책은 과거에는 존재하지 않았던 수요에 대한 대응이라는 해석보다는 기존의 아날로그 자료가 전산화되고 수많은 자료들로 인해 새롭게 생성되는 빅데이터라는 대상과 이를 분석할 수 있는 인공지능(AI)이라는 방법이 등장하였기 때문이라 볼 수 있다. 또한, 이러한 시대적 흐름에 맞춰 국군도 2020년의 주요 업무 중 하나로 4차 산업혁명 기술을 적용하는 것을 목표로 하고 이를 위한 빅데이터와 인공지능의 적용을 시도하고 있다(국방부 2020).

빅데이터는 현대에 들어 그 중요성이 강조되고 있지만, 빅데이터라는 수집된 대량의 데이터 그 자체보다는 데이터를 발굴・정리・분석하는 등 이를 어떻게 활용하느냐에 따라 진가를 발휘한다. 이처럼 정보 과학(Information Science) 분야에서 데이터는 원재료로서 분석과 가공을 통해서 정보(Information)가 되고 이러한 정보가 사용자에게 적절하게 사용되었을 때 지식(Knowledge)으로 변환됨으로써 도움이 된다고 보았다(Zins 2007). 이와 같은 점을 고려할 때 국방 분야에서의 효과적인 지식의 생산을 위해서는 지식의 원재료인 데이터에 대한 이해와 이를 적절히 가공해내는 방법에 대한 고려가 필요하다.

본 연구에서는 많은 데이터의 형태 중 한국인이 일상적으로 자연스럽게 사용하는 언어 형태인 한글 형태의 자연어를 국방 분야에 최적화하여 분석하는 방법을 확인하고자 하였다. 이는, 비정형화된 데이터인 문자, 이미지, 음성, 영상 등이 빅데이터의 다량을 차지하고있고, 그 중에서도 한글로 이루어진 각종 국방 분야의 문서들의 경우 텍스트 마이닝 기법을 통해 여러 연구에서 분석이 활발히 이루어지고 있기 때문이다. 따라서, 여러 연구들에서 이러한 비정형 문서들을 자연어 처리(Natural Language Process)를 통해 연구자와 컴퓨터가 처리할 수 있는 형태로 변형 후 이를 분석하였다. 국방 분야의 문서를 자연어 처리를 통해 연구를 수행할 경우 유의해야 할 점 중 하나는 국방의 고유한 용어들에 대한 정확도이다. 자연어 처리를 위한 많은 상용 형태소

분석기의 경우 일반적인 단어사전이나 트위터, 기사 등의 자료를 기반으로 학습되었다. 이러한 형태소 분석기를 사용하여 분석을 진행할 경우 국방관련 용어를 정확히 반영하지 못함으로써 자료가 누락되거나 일부 단어에 편향(Bias)된 자료를 사용함으로써 분석 자체의 정확성을 낮출 우려가 있다.

따라서 본 연구에서는 상용 형태소 분석기와 국방 및 군사 분야의 사전을 동시에 활용함으로써 비교적 빠르고 간편하면서도 국방 분야 빅데이터 자연어 분석에서 정확성을 높이는 방식을 제시함으로써 효율적이면서도 효과적인 자연어 분석 방법을 제시하고자 한다.

Ⅱ. 선행연구 고찰

1. 형태소 분석기의 종류와 특징

형태소란 의미를 갖는 최소 단위의 언어이자 형식의 최소 단위이다(이선웅 외 2017). 형태소 분석은 주어진 문장을 이러한 형태소 단위로 분석하는 것으로, 분석을 통해 얻어낸 각각의 형태소들의 의미를 바탕으로 문장과 문장이 담긴 문서의 의미를 추론할 수 있다. 형태소 분석은 비정형 문자의 분석을 위해서는 필수적이다. 이는, 문장을 바로 분석할 경우 기계는 '국방을', '국방이', '국방에', '국방과' 같이 조사만 달라지더라도 이를 별개로 판단하기 때문이다. 이를 형태소 분석을 통해 '국방' + '-을/-이/-에/-과'로 분석한다면 해당 문장들이 '국방'과 관련한 내용을 포함하고 있음을 알 수 있게된다.

언어에 따라 문자나 문법 등이 달라질 수 있으므로 일반적으로 형태소 분석기는 분석 대상 언어에 따라서 달라진다. 특히, 한국어의 경우 다른 언어와 달리 한글이라는 고유의 문자체계를 사용하므로 이에 맞는 형태소 분석기를 사용한다면 분석 성능을 향상시킬 가능성이 높아질 것이다. 이러한 관점에서 <표 1>과 같이 연구자가 쉽게 사용할 수 있는 다양한 방식의 오픈소스 한국어 형태소 분석기가 개발되었다. 이들 형태소 분석기는 각각 비지도 기계학습 및 분류를 위해 CRFs(Conditional Random Fields), HMM(Hidden Markov Models), CNN(Convolutional Neural Network)를 위주로 사용하되 각각 분류성능 향상 및 최적화를 위해 모델을 세부화하고 파라메터를 조정하였다.

<표 1> 한글 형태소 분석기의 종류

분석기 / 개발연도	주요 알고리즘	말뭉치(Corpus)
Hannanum / 1999 (한국과학기술원 2011)	HMM	KAIST Corpus
KKMA / 2010 (이동주 외 2010)	HMM	세종
Komoran / 2013 (신준수 외 2016)	HMM	세종
Mecab-ko / 2013 (Kudo 외 2004)	CRFs	세종
Khaiii / 2018 (임재수 2018)	CNN	세종

많은 형태소 분석기의 알고리즘이 학습된 자료에 기반하였다는 점을 고려할 때 형태소 분석기의 성능을 좌우하는 요인 중 하나는 학습에 사용된 데이터인 말뭉치이다. 많은 오픈소스 형태소 분석기는 21세기 세종계획에 따른 결과물인 세종 말뭉치를 사용하고 있다. 세종 말뭉치는 2007년 이전까지의 다양한 신문, 잡지, 도서와 같은 문어와 방송, 녹음자료 등을 활용한 구어 자료를 활용하여 구축된 자료이다(국립국어원 2007; 황용주 외 2016). 하지만, 말뭉치에 의해서 학습된 형태소 분석기는 말뭉치 자체에 오류가 존재하거나 말뭉치에 없었던 신조어나 특정 전문분야의 단어가 출현할 경우 분석 성능이 낮아진다는 단점이 존재하며(김노은 외 2019; 최민석 외 2020) 세종 말뭉치 또한 신문, 잡지, 도서 등의 자료로 국방 분야에 대한 충분한 자료를 수집하기 어려웠을 것으로 추정된다. 따라서 상용 형태소 분석기를 사용하여 국방 분야의 비정형 한글 데이터를 다룰 때 정확도를 높이기 위해서는 국방 관련 단어를 추가로 분석기에 학습시키거나 별도로 분류할 수 있도록 사용자 사전의 형태로 제공되어야 할 것이다.

2. 국방관련 연구에서의 자연어 처리

최근 국방 분야에서도 다양한 연구들이 연구 과정에서 형태소 분석을 사용함으로써 비정형 데이터를 정형화한 후 해당 자료를 바탕으로 연구를 수행하였다. 국방 분야에서의 형태소 분석기를 활용한 연구들은 주로 대량의 문서들을 형태소 분석기를 사용

하여 정형화하고 정형화된 자료를 바탕으로 주제를 도출하기 위한 방법을 사용하였다. 주제 도출은 간단하게는 단어 빈도 분석(Term Frequency, TF) 및 빈도-역문서빈도(Term Frequency-Inverse Term Frequency, TF-IDF)를 활용한 분석(서호준 2019; 임상수 외 2019)에서부터 문장 내 출현 단어나 문서 내 유사도를 통해 단어·문서 간의 연결을 분석하는 네트워크 분석(Network Analysis) 방법(이용규 외 2018; 서호준 2019; 임상수 외 2019) 그리고 문서 내의 단어의 출현 확률을 통해 전체 문서 집합의 주제 및 해당 주제의 단어를 추론하는 잠재 디리클레 할당(Latent Dirichlet Allocation) 등의 분석 방법(임상수 외 2019; 이동혁 외 2020; 전고운 외 2020)이 주로 사용되었다.

위와 같은 분석 방법들은 모두 전처리의 단계에서 최소 의미를 지닌 단어를 어떻게 분류하느냐가 주요하게 작용한다. 예를 들어 '국방표준화정보체계'라는 단어를 형태소 분석기가 '국방 + 표준 + 화 + 정보 + 체계"와 같이 분석한다면 이들 단어는 각각 (국방, w_1), (표준, w_2), (화, w_3), (정보, w_4), (체계, w_5)의 다른 형태로 인식하게 된다. 이렇듯 특정 단어를 제대로 인식하지 못할 경우 빈도분석에서는 특정 단어의 빈도가, 네트워크 분석이나 토픽 모델링에서는 특정 단어나 특정 문서들의 연관성이 과소 혹은 과대하게 분석되는 문제가 발생하게 된다. 그러므로 국방과 같이 특정 분야의 비정형 문자 데이터를 사용하여 연구할 경우 국방에 맞는 형태소 분석을 고려한 연구가 수행될 필요성이 있다.

최근에 비정형 문자 데이터를 사용하여 수행된 <표 2>의 국방 분야의 연구들의 특징을 살펴보면 국방 관련 문서 처리를 위한 자연어 처리 사전 구축이 필요함을 알 수 있다. 첫째로, 관련 연구들의 분석 대상이 비교적 학습이 잘 되어있는 신문기사나 기고문에서부터 국방 분야의 전문 용어가 사용되는 학술자료 및 공문서까지 다양하게 분포되어있다. 신문기사의 경우 새롭게 등장한 국방관련 단어가 아닐 경우를 제외하면 오분류의 가능성이 비교적 낮지만, 학술자료나 공문서의 경우 국방관련 용어가 자주 사용되므로 국방관련 사전의 필요성이 높다. 둘째로, 사용된 형태소 분석기가 주로 머신러닝에 기반하였다는 점이다. 형태소 분석기가 머신러닝이 아닌 특정한 문법 규칙을 기준으로 분류할 경우 문서가 단순한 언어 규칙을 가지고 있을 경우 분류의 정확도가 높아지나 다양한 데이터로부터 분석을 수행할 경우 정확도가 낮아진다는 단점이 존재한다(임해창 외 1994; 이재성 2011; 서민영 외 2018). 따라서 국방분야의 연구에서는 주로 다양한 데이터에 대한 분류를 위해 개발된 머신러닝에 기반한 형태소 분

석기를 사용하였다. 하지만 이러한 형태소 분석기들의 경우 학습데이터를 기반으로 분류하므로 이를 통해 만들어진 학습 사전에 없는 단어들을 처리하는 것이 어렵다는 단점이 존재한다(노태길 외 2000; 임좌상 외 2014). 마지막으로 앞선 특정 분야의 연구에서 머신러닝의 단점을 극복하기 위한 사용자 사전의 활용이 미비하였다. 연구자들은 국방분야의 용어를 별도로 학습시키지 않거나, 연구자의 목적에 따라 특정 단어만 수정하여 사용하는 방식을 사용하였다. 전자의 경우 형태소 분석기의 과학습(overfitting)된 자료에 의한 편향이 발생하며, 후자의 경우 연구자의 목적에 따른 단어들만 과분류 될 수 있다는 위험이 존재한다. 앞선 국방분야 연구에서 형태소 분석의 특징을 고려할 때 국방분야 데이터 분석을 위한 자연어 사전은 연구의 편향성을 낮출 수 있을 것이다.

<표 2> 형태소 분석을 이용한 국방 분야 연구

연구자(연도)	분석대상	형태소 분석기	사용자 사전 활용여부
이용규 외(2018)	논문초록	NetMiner[2]	X (사용여부 미명시)
백승원 외(2018)	신문기사	미명시 (KoNLPy 사용)	X (사용여부 미명시)
서호준(2019)	정부문서 (국방백서)	Espresso K (Textom[3] 사용)	X (사용여부 미명시)
임상수 외(2019)	정부문서 (군내부공문)	Hannanum (KoNLP 사용)	X (사용여부 미명시)
이동혁 외(2020)	기고문	Komoran	연구자 자체 단어 수정
전고운 외(2020)	DTiMS 제공 영문기사 (국문번역)	Hannanum (KoNLP 사용)	연구자 자체 단어 수정

2) NetMiner는 네트워크 분석을 위한 상용프로그램으로 자체 형태소 분석기능을 제공한다.
3) Textom은 빅데이터 분석을 위한 상용프로그램으로 Mecab-Ko와 Espresso K 형태소 분석기능을 탑재하고 있으며, 해당 연구에서 사용된 Espresso K는 HMM에 기반한 형태소 분석기이다(홍진표 2009).

Ⅲ. 국방분야 자연어 사전 구축

1. 자료 선정 및 수집

본 연구는 두 가지 부분에 초점을 두고 국방분야 비정형 데이터 처리를 위한 사전 구축하고자 하였다. 첫째는 자료 사용에 소요되는 시간 및 비용을 최소화하고자 하였다. 국방분야 빅데이터의 활발한 분석을 위해서는 별도의 비용이 들지 않으면서도 손쉽게 활용할 수 있는 자료를 사용하여야 한다. 인터넷에 공개된 국방 및 군사 관련 용어를 활용할 경우 별도의 비용이 들지 않고, 빠르게 국방에 관한 비정형 데이터 처리를 위한 광범위한 사전 데이터를 구축할 수 있으므로 이를 활용하였다. 둘째는 기존의 형태소 분석기의 말뭉치에 사용되지 않은 자료이되, 신뢰성 있는 자료를 사용하고자 하였다. 기존의 다양한 형태소 분석기는 기사나 도서에서부터 트위터와 같은 SNS까지 다양한 비정형 데이터를 학습자료로 활용하였으나, 국방과 같은 특정 전문분야의 용어는 학습하지 못하였다는 단점을 가지고 있다. 이러한 특징을 고려할 때 국방분야의 전문용어를 반영할 수 있는 신뢰성있는 자료를 사용할 경우 앞선 형태소 분석기들의 단점을 보완할 수 있을 것이다. 국방 및 군사와 관련된 사전은 이러한 요건을 충족할 수 있는 자료이다. 해당 분야의 전문가에 의해서 구축된 전문용어사전은 개별 의미를 가지는 각각의 용어를 기술하였으므로 지도학습자료로 활용할 수 있다.

앞선 두 사항을 모두 고려하여 연구에서는 인터넷에 공개된 국방 및 군사용어 관련 사전을 사용하고자 하였다. 연구에 사용된 사전은 ㈜네이버에서 제공[4]하는 국방과학기술용어사전, 군사용어사전, 병역관련 용어해설과 전쟁기념관에서 제공하는 전쟁군사용어 사전[5], 그리고 합동참모본부의 군사용어사전[6]의 총 5개 사전을 활용하였다. 각 사전들은 용어를 한글・영문, 혹은 한글・영문・한자로 제공하는 사전으로 본 연구에서는 한글로 된 비정형 데이터를 처리하는 것에 초점을 맞추었으므로 한글 자료를 위주로 사용하되 한자의 경우 음을 한글로 변환하여 사용하였다.

---

4) ㈜네이버는 네이버 지식백과(https://terms.naver.com/list.nhn?cid=42156&categoryId=42156)에서 국방기술품질원(국방과학기술용어사전), 군사용어사전(이태규, 일월서각), 병역관련용어해설(병무청) 자료를 제공하고 있음.

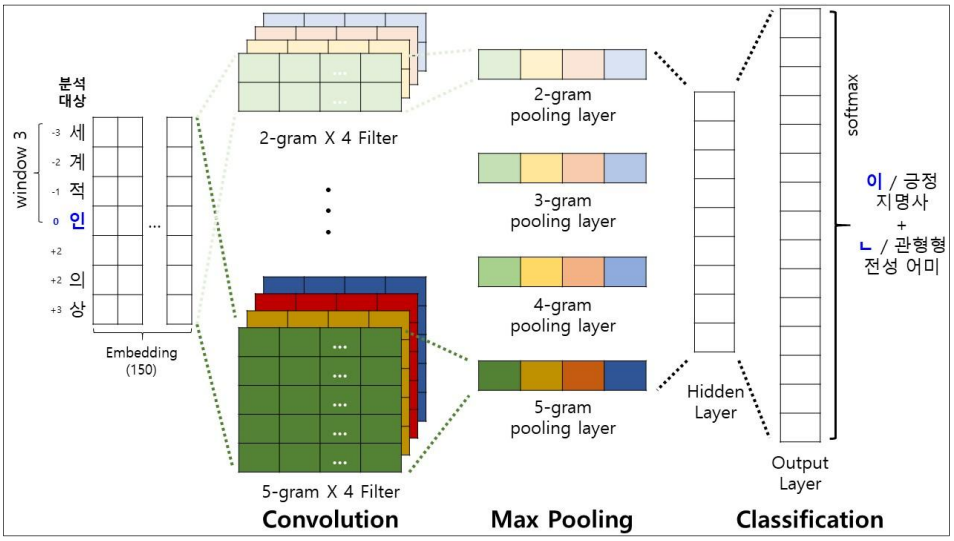
5) 전쟁기념사업회는 전쟁・군사・유물 정보(https://www.warmemo.or.kr/front/militaryInfo/searchList.do)에서 전쟁군사 용어사전을 한・영으로 제공하고 있음.

6) 합동참모본부는 군사용어해설(https://www.jcs.mil.kr/user/indexSub.action?codyMenuSeq=71157&siteId=jcs&menuUIType=sub)에서 합동참고교범 및 합도연합작전 군사용어사전을 참고한 한글용어사전을 제공하고 있음.

사전 자료 수집을 위해 Python 3.8 소프트웨어에서 Selenium 라이브러리를 활용한 ChromeDriver 88.0로 단어를 크롤링(crawling)하였다. 개별 사전으로부터 수집된 용어들은 공백, 특수문자. 숫자, 영어를 제거하는 전처리 과정을 거친 뒤 한글자 단어와 각 사전 간의 중복단어를 제외하여 총 20,785개의 단어로 구성된 국방분야 자연어 사전을 구축하였다.

2. 형태소 분석기와의 결합 및 비교 방안

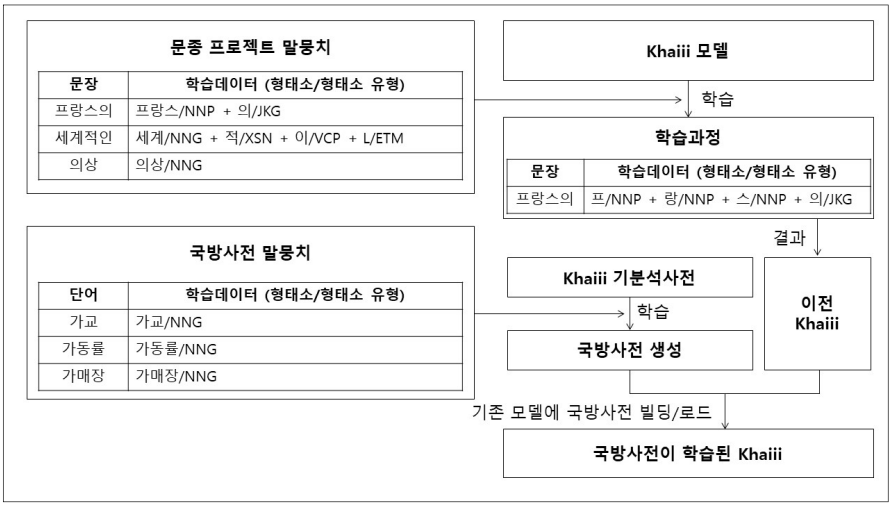
국방분야 자연어 사전의 활용을 위해서 연구는 기존에 사용되는 형태소 분석기에 국방분야 자연어 사전을 학습시키는 방식을 사용하고자 하였다. 학습에 사용된 형태소 분석기는 주식회사 카카오에서 개발된 Khaiii로 다른 형태소 분석기보다 비교적 최근에 개발되었으며 '문종 프로젝트'라는 작업을 통해 세종 말뭉치의 오류를 개선하여 성능을 높이고자 하였고, 사용자가 자체 사전파일을 학습시킬 수 있는 기분석 사전 기능을 보유 하고있어 국방분야 자연어 사전의 활용이 용이하다는 장점이 있다. 따라서 본 연구에서는 국방분야 자연어 사전의 20,785개 단어 중 Khaiii의 이전 학습 자료인 문종 프로젝트 말뭉치와 충돌하는 687개의 단어를 제외한 20,098개의 단어를 학습에 사용하였다.



<그림 1> Khaiii의 CNN 모델 네트워크 구조

Khaiii는 <그림 1>과 같은 CNN 모델 구조에서 문종 프로젝트를 통해 만들어진 말뭉치를 학습하여 문장의 형태소를 분석한다(임재수 2018). Khaiii는 분석 음절의 앞·뒤 3개 음절을 각각 2~5개의 단위로 묶어 4개의 Convolution 필터를 거치게 하며 이를 통해 산출된 벡터에서 Max-over-time pooling을 통해 각 레이어별 최대값을 산출한다(Kim 2014; 임재수 2018). 이후 2~5의 커널 크기에서 4개의 필터별로 나온 최대값을 연결하고 연결된 하나의 백터가 잠재 레이어와 출력 레이어를 통과하여 음절의 형태소를 결정하게 된다(임재수 2018).

국방분야 용어에 대한 분석성능을 비교하기 위하여 연구는 실험군인 국방사전이 학습된 Khaiii 형태소 분석기와 사전을 학습하지 않은 통제군인 이전 Khaiii, Hannnanum, Kkma, Mecab-ko, Komoran의 5개 형태소 분석기를 사용하였다.



<그림 2> Khaiii의 국방사전 학습 과정

실험군인 국방사전이 학습된 Khaiii의 학습과정은 <그림 2>와 같다. 연구는 Khaiii 모델의 분석성능을 향상시키기 위해 Khaiii 모델의 기분석 사전 기능을 활용하여 국방사전을 모델에 학습시키고자 하였다. Khaiii 모델의 기분석 사전 기능은 연구자가 원하는 단어나 문장 말뭉치와 해당 말뭉치의 형태소 유형을 학습시키고 이를 Khaiii가 지닌 기존 학습모델에 불러올 경우 형태소 분석에서 사전의 내용을 반영시키는 방법으로 4개 음절 이상의 단어 분석에서 분석률을 향상시키기에 용이하다(임재수 2018). 따라서, 앞서 구축한 국방사전의 20,098개 단어를 Khaiii 용 기분석 사전으로

만들고 이를 기존의 Khaiii 모델에 설치하여 국방사전이 학습된 Khaiii를 구축하였다.[7)]

분석 성능 비교를 위한 분석 대상은 국방사전 구축을 위해서 수집한 국방분야 용어로 20,785개의 명사로 구성되어있다. 해당 분석대상에서 최소음절을 가진 단어는 2음절, 최대음절을 가진 단어는 19음절로 나타났으며 단어들의 평균음절은 약 5.48음절이다. 연구는 각 형태소 분석기가 분석 대상을 형태소 분석하여 이를 1개의 고유한 단어로 인식한 단어의 수를 측정하는 방식으로 수행하였다. 예를 들어, '국방표준화정보체계'라는 1개의 단어를 각 형태소 분석기에서 분석하도록 한 뒤, 분석결과가 '국방' + '표준' + '화' + '정보체계'와 같이 원래의 단어와 다른 형태로 분석한 경우를 모두 제외하고, 정확히 '이동차단작전'이라는 1개의 단어로 인식할 경우 정확히 분석한 것으로 보았다. 이는 문장 내에서 해당 단어를 제대로 인식하는지를 측정하는 방법에 비해서는 연구의 강건성이 떨어지나 성능 비교를 위한 별도의 정답문헌을 생성하지 않아도 되므로 성능비교에 사용되는 시간과 비용을 절약할 수 있고, 고유한 용어를 분해하여 각각의 다른 단어로 인식하거나 조사 등으로 인식하여 제거하는 오류를 발견할 수 있다는 점에서 적절한 연구 방식이다.

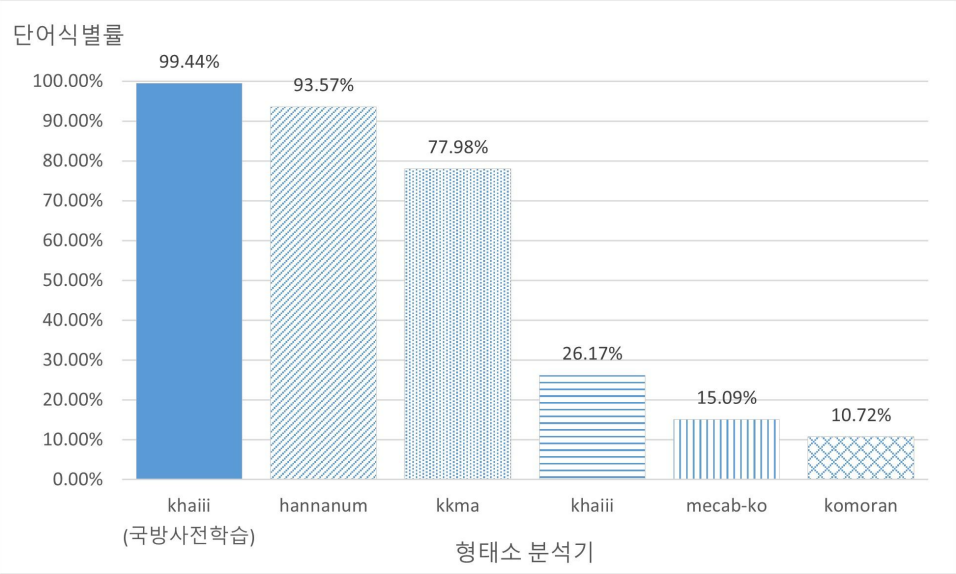
각각의 형태소 분석기는 해당 분석기의 동작가능 환경을 고려하여 국방사전이 학습된 Khaiii, 이전 Khaiii, hannanum, mecab-ko는 Ubuntu 18.04 OS의 Python 3.8 버전에서. Kkma와 Komoran은 Windows 10 OS의 Python 3.8 버전에서 분석을 수행하였다.

Ⅳ. 자연어 사전 구축 결과 분석

형태소 분석기 별 국방분야 용어에 대한 분석 성능을 비교한 결과는 <그림 3>과 같다. 분석 결과 국방사전이 학습된 Khaiii 형태소 분석기의 국방관련 단어식별률이 99.43%로 측정되어 가장 높은 분석 성능을 보였으며, Hannanum, Kkma, 이전 Khaiii, mecab-ko, komoran의 순으로 국방분야 용어를 식별하였음을 알 수 있다. 하지만

7) AMD Ryzen 7 4700U 프로세서로 Windows 10의 WSL을 이용한 Ubuntu 환경에서 약 2만 단어의 기분석 사전 빌딩 및 로드에는 약 60분 가량이 소요되었으며, 사전을 학습 이후에는 국방사전을 활용하기 위해 필요한 별도의 시간 소요는 없었다. 이러한 학습에 필요한 시간은 운영체제 및 컴퓨터 환경 등에 따라서 차이가 있을 수 있다.

Hannanum 형태소 분석기의 경우 분석 결과가 우수하게 측정된 이유는 실제 분석 성능의 우수성보다는 형태소 분석기의 특성과 본 연구 방식의 특성으로 인해 오판된 것일 가능성이 높다. 이는 Hannanum 형태소 분석기는 띄어쓰기가 없는 단어를 새로운 형태소로 인식하는 경향이 있으므로(김은주 외 2014) 띄어쓰기가 없는 개별 단어를 대상으로 분석할 경우 이를 하나의 새로운 단어로 인식하기 때문에 인식률이 높은 것으로 보일 수 있다. 따라서 실제 문장 내에서 국방관련 용어를 식별할 경우 이보다 낮은 성능을 보일 것으로 추정된다.



<그림 3> 국방사전 학습 여부에 따른 형태소 분석기 비교

연구의 결과는 두 가지 측면에서 비정형 데이터 분석에서 국방관련 자연어 사전을 활용하는 것이 필요함을 보여준다. 첫째로 국방사전을 학습한 형태소 분석기가 다른 형태소 분석기와 비교할 때 국방분야의 용어 분류에서 상대적으로 높은 성능을 보인다는 점이다. 비정형 문자 데이터의 분석에 있어서 적절한 형태소 분석기의 선정과 최적화는 연구의 방법론과 무관하게 연구 결과를 결정짓는 중요한 요소 중 하나이다. 국방분야와 관련된 비정형 데이터의 분석에서 연구와 같은 국방사전을 학습한 형태소 분석기의 사용은 연구에 사용되는 데이터의 정형화 과정에서 발생할 수 있는 국방용어에 대한 편향을 낮춤으로써 분석결과의 신뢰도를 높일 수 있다. 둘째로는 동일 형

태소 분석기에서 국방사전의 학습 여부가 국방용어에 대한 식별률을 높일 수 있다는 점이다. 연구에서는 다양한 형태소 분석기 중 Khaiii를 실험군으로 국방사전을 학습시켜 국방용어 식별률을 확인하였다. 연구 결과는 통제군인 국방사전을 학습하지 않은 이전 Khaiii는 26.17%의 식별률을, 실험군인 국방사전이 학습된 Khaiii는 99.44%의 식별률을 보여주었다. 이러한 결과는 동일한 형태소 분석기를 사용하더라도 국방분야에 대한 전문용어의 학습여부가 정확한 비정형 문자 데이터의 분류에 큰 영향을 미칠 수 있음을 시사한다. 결론적으로 위와 같은 사유들은 국방에 관한 비정형 문자 데이터의 분석에서 국방용어사전을 형태소 분석기에 학습시키는 것이 연구의 편향을 낮추고 연구 결과의 신뢰성을 높일 수 있는 방안이 될 수 있음을 보여준다.

Ⅴ. 결 론

국방분야에서 빅데이터를 활용한 분석에 대한 관심은 커져가고 있으며, 특히 한글을 기반으로 한 비정형 데이터를 정형화하는 과정을 통해 연구나 정책과 관련한 자료를 수집하고 이를 분석하고자 하는 시도가 증가하고 있다. 본 연구는 이러한 흐름에 맞춰 비정형 한글 데이터를 정형화하는 과정 중에서도 형태소 분석기의 학습 데이터에 의해 생겨날 수 있는 편향을 낮추기 위한 방법으로 국방용어에 특화된 자연어 처리 사전의 활용을 제시하였다. 연구에서는 인터넷에 공개된 5개의 국방/군사와 관련한 사전을 수집 및 정제하여 낮은 비용 및 짧은 시간 내에 신뢰성 있는 국방분야 자연어 처리 사전을 구축하였다. 구축된 사전을 형태소 분석기에 학습시켜 기존의 오픈소스 형태소 분석기들과 비교한 결과 국방용어의 분류에 있어 연구에서 제시한 국방분야 자연어 처리 사전을 학습한 모델이 다른 모델들에 비해 우수한 성능을 나타내었다.

연구의 결과는 다음과 같은 시사점을 제시한다. 첫째로, 머신러닝에 기반한 형태소 분석기를 사용할 경우 해당 분석기의 학습 자료를 고려하여 신중한 사용이 필요하다. 다수의 머신러닝 기반 형태소 분석기의 경우 다양한 자료에서의 범용성을 위해 기사나 도서, 인터넷 게시물들을 사용하고 있다. 이러한 형태소 분석기를 사용하여 국방에 관한 전문적인 용어가 사용된 글을 분석할 경우 고유한 의미를 지닌 하나의 단어가 분해되어 여러 단어로 나뉘거나 잘못된 형태로 식별될 수 있다. 연구 결과 또한 형태소 분석기에 따라 단어 식별률에서 큰 차이를 보였다. 따라서 학습자료에 대한 의존

성이 있는 형태소 분석기를 사용하여 국방관련 비정형 데이터에 대해 분석할 경우 해당 자료의 대표적인 표본을 선정하고 여러 형태소 분석기를 사용하여 표본에 대한 분석력이 높은 형태소 분석기를 선택하는 등의 과정이 필요하다.

두 번째로, 국방분야 비정형 데이터의 원활한 연구와 분석을 위해서는 국방용어사전 및 학습 데이터가 구축되어야 한다. 국방과 관련한 비정형 데이터를 사용한 이전의 몇몇 연구들은 별도의 사전을 사용하지 않거나 연구자가 지정한 특정 단어만 별도로 처리함으로써 편향된 연구 결과가 도출될 가능성이 잔재하였다. 연구의 결과 또한 사전의 학습 여부가 형태소 분석기의 성능에 영향을 미칠 수 있음을 보여주었다. 하지만 연구에 사용된 사전 또한 크롤링을 통해 수집한 약 2만 단어의 말뭉치로 데이터의 양이 한정적이며, 저작권의 문제로 인해 재배포나 다른 연구자와의 협업을 통한 오류 수정이 어렵다는 단점이 존재한다. 따라서 여러 연구자가 동일한 환경에서 분석할 수 있고, 분석된 결과의 정확성을 높이기 위해서는 기준이 될 수 있는 국방용어사전 및 국방관련 학습 데이터가 필요하다.

끝으로 연구는 머신러닝 기반의 오픈소스 형태소 분석기를 사용할 때 효율적으로 국방분야의 비정형 한글 데이터를 분석할 수 있는 방법으로 국방분야 사전 학습을 제시하였다. 하지만 이러한 방법은 여러 한계도 지니고 있다. 연구에서 국방분야 자연어 사전 구축을 위해 인터넷 자료를 크롤링하여 불용어가 포함되어있을 수 있고 띄어쓰기가 있는 단어의 경우 이를 제거하여 한 단어로 처리하는 방식을 사용하였다. 이러한 구축 방식으로 인해 연구에서 사용된 국방분야 자연어 사전은 분석 성능에 한계가 있다. 또한, 자연어 사전 구축 후 이를 형태소 분석기에 학습시키는 방식은 규칙 기반의 형태소 분석기에는 적용하기 어렵다는 단점이 존재한다. 따라서 후속 연구에서는 앞선 한계들을 고려하여 다양한 학습 어휘를 구비하고 검증할 수 있는 방법과 이를 다양한 형태소 분석기에 적용할 수 있는 방안을 고려해야 할 것이다.

< 참 고 문 헌 >

1. 국립국어원. 2007. 「21세기 세종계획 국어 기초자료 구축」. 서울 : 현대문화사.
2. 국방부. 2020. “2020년 국방부 업무보고 : 국민과 함께 평화를 만드는 강한 국방.” https://www.mnd.go.kr/mbshome/mbs/plan/download/plan.pdf (작성일 : 2020.01.21) (검색일 : 2021.01.02.).
3. 김노은·정상근. 2019. “데이터 기반 형태소 분석기의 신규명사 추출 경향 분석.” 「한국정보과학회 학술발표논문집」 1421-1423.
4. 김은주·최정우·류정우. 2014. “형태소 분석을 이용한 발화관련 기기의 새로운 입력 키워드 추출.” 「한국화재소방학회 논문지」28(2).
5. 노태길·이상조. 2000. “규칙 기반의 기계학습을 통한 고유명사의 추출과 분류.” 「한국정보과학회 학술발표논문집」 27(2), 170-172.
6. 백승원·한승헌·이창준·이지섭·문수환. 2018. “비정형데이터 기반 공공 건설사업 갈등 이슈 분석: 제주해군기지 사례를 중심으로.” 「공공사회연구」 8(1), 83-106.
7. 서민영·홍태성·김주애·고영중·서정연. 2018. “계층 구조 어텐션 매커니즘에 기반한 CNN-RNN을 이용한 한국어 화행 분석 시스템.” 「한국정보과학회 언어공학연구회:학술대회논문집」, 243-246.
8. 서호준. 2019. “우리나라 국방정책의 핵심이슈 도출-[2018 국방백서] 에 대한 텍스트 네트워크 분석의 적용.” 「한국군사」6, 39-70.
9. 신준수·이승원. 2016. “KOMORAN 3.0” 2016 국어 정보 처리 시스템 경진대회 발표 자료집. https://korean.go.kr/common/download.do;front=57EB12EDEFEA255BB842C7D9697209C0?file_path=etcData&c_file_name=c00c0198-220a-47a5-bccd-55eda862dbe4_0.pdf&o_file_name=2016년 국어 처리 정보 시스템 경진 대회 발표 자료집.pdf
10. 이동주·연종흠·황인범·이상구. 2010. “꼬꼬마: 관계형 데이터베이스를 활용한 세종 말뭉치 활용 도구.” 「정보과학회논문지: 컴퓨팅의 실제 및 레터」 16(11), 1046-1050.
11. 이동혁·이서희. 2020. “정신전력 기고문의 토픽모델링 적용 탐색적 연구-국방일보 오피니언 분석을 중심으로.” 「정신전력연구」 61, 103-133.
12. 이선웅·오규환. 2017. “형태소의 식별과 분류.” 「국어학 (國語學)」 81, 263-

294.

13. 이용규·윤성웅·이상훈. 2018. “텍스트 네트워크분석을 활용한 국방분야 연구논문 지식구조 분석.” 「한국컴퓨터정보학회 학술발표논문집」 26(2), 525-528.
14. 이재성. 2011. “한국어 형태소 분석을 위한 3단계 확률 모델.”「정보과학회논문지 : 소프트웨어 및 응용」 38(5), 257-268.
15. 임상수·이문걸. 2019. “군 조직 업무 분석기법에 관한 연구.” 「한국데이터정보과학회지」30(1), 139-157.
16. 임좌상·김진만. 2014. “한국어 트위터의 감정 분류를 위한 기계학습의 실증적 비교.” 「한국멀티미디어학회논문지」17(2), 232-239.
17. 임재수. 2018. “카카오의 딥러닝 기반 형태소 분석기.” https://brunch.co.kr/@kakao-it/308. (검색일 : 2021.01.04.)
18. 임해창·임희석·윤보현. 1994. “[기술해설] 자연어처리 연구동향: 통계 기반의 자연어 처리.” 「정보과학회지」 12(9), 20-30.
19. 전고운·강인원·전정환. 2020. “토픽모델링 기반의 국방기술 동향분석 방안: 장갑전투차량에의 적용.” 「산업혁신연구」 36(1), 69-94.
20. 최민석·김창현·박호민·천민아·윤호·남궁영·김재균·김재훈. 2020. “XGBoost 와 교차검증을 이용한 품사부착말뭉치에서의 오류 탐지.” 「정보처리학회논문지/소프트웨어 및 데이터 공학」 제9(7), 7.
21. 한국과학기술원. 2011. 「한나눔 한국어 형태소 분석기 사용자 매뉴얼.」 대전 : 시멘틱 웹 첨단연구센터, 한국과학기술원.
22. 홍진표. 2009. “어절패턴 사전을 이용한 한국어 품사 태거.” 창원대학교 컴퓨터공학과 석사학위논문.
23. 황용주·최정도. 2016. “21 세기 세종 말뭉치 제대로 살펴보기—언어정보나눔터 활용하기.” 「새국어생활」26(2), 0-0.
24. Yoon Kim. 2014. “Convolutional Neural Networks for Sentence Classification.” *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1746-1751.
25. Kudo, T., Yamamoto, K., & Matsumoto, Y. 2004. “Applying conditional random fields to Japanese morphological analysis.” *In Proceedings of the 2004 conference on empirical methods in natural language processing*, 230-237.

26. Zins, C. 2007. “Conceptual approaches for defining data, information, and knowledge”. *Journal of the American society for information science and technology* 58(4), 479-493.

원고 접수: 2021. 03. 05. / 수정 접수: 2021. 04. 27. / 게재확정: 2021. 04. 29.